基于核心主题特征的作者身份识别研究

孟 旭 [1,2]**, 谢 靖** [1,*]**, 李春旺** [3]

1 中国科学院文献情报中心

2 中国科学院大学经济与管理学院图书情报与档案管理系

3 中国科学院计算机技术研究所

摘要：【背景及目的】作者识别正在向多层次特征的使用发展，而相较于文体风格特征，主题特征在历来作者识别研究应用中仍是少数，特别是针对中文社交媒体文本的作者识别。同时针对主题特征的利用研究，更多的是对主题特征的抽取技术和方法的创新，而未对识别出的主题以及主题特征的应用方法进行进一步研究。所以，本研究以主题特征在中文社交媒体文本作者识别中的使用研究为基本目的，同时进一步制定策略对主题特征中的核心主题进行识别和筛选，优化主题特征的使用方法，从而提高主题特征在作者识别中的使用效果。【方法】研究首先利用 LDA 主题模型抽取候选作者的学术主题和社交主题，然后利用 word2vec 制定合并筛选策略进行核心主题的识别和表示，最后结合 N-gram 特征和相似度计算的办法实现作者识别。【结果】实验结果显示主题特征在本研究语料上对作者识别有一定的积极作用，同时本研究提出的核心主题特征相关策略和应用也能优化主题特征的使用效果。

关键词：作者身份识别；主题特征；N-gram；科研作者；社交网络文本

分类号：G206

Research on author attribution based on core topic

MENG Xu, XIE Jing, LI Chunwang

Abstract: [Background and purpose] Author recognition is developing towards the use of multi-level features. Compared with stylistic features, thematic features are still a few in the research and application of author recognition, especially for Chinese social media texts. At the same time, the research on the use of topic features focuses more on the innovation of the extraction technology and methods of topic features, but not on the identified topics and the application methods of topic features. Therefore, the basic purpose of this study is to study the use of topic features in the author recognition of Chinese social media texts, and further develop strategies to identify and screen the core topics in the topic features, optimize the use of topic features, so as to improve the use effect of topic features in the author recognition. [Methods] The research first uses the LDA topic model to extract the academic topics and social topics of the candidate authors, and then uses Word2vec to develop a merge screening strategy to identify and represent the core topics, and finally uses N-gram features and similarity calculation to achieve author recognition. [Results] The experimental results showed that the thematic features had a certain

---

* 通讯作者：谢靖，xiej@mail.las.ac.cn

positive effect on the author's recognition in the corpus of this study, and the strategies and applications related to the core thematic features proposed in this study could also optimize the use of thematic features.

Key words: Author attribution; Topic characteristics; N-gram; Scientific research author; Social media text

Class Number：G206

0 引言

近年来，关于中文社交媒体文本作者身份识别研究一直受到关注，已经取得一些研究成果，这些研究主要实现不同网络平台、不同社区、不同话题中同一作者信息的识别，识别方法主要基于文体风格特征，而利用文本主题特征研究不够。同时，针对科研人员的人才评价等工作随着进入大数据时代，不仅要利用学术论文等数据，还要利用科研人员相关的社交媒体信息、学术交流信息、教学信息等多类型数据，这使得针对科研人员的信息集成也成为关注方向。

本研究将基于存在的类似科学网的实名认证社交媒体平台，在作者识别任务已有研究的基础上，重点研究主题特征在中文社交媒体文本作者识别中的作用和应用意义，同时结合科研人员发表的学术论文信息制定筛选核心主题特征的相关策略提取作者的核心主题特征，并考察构建的核心主题特征是否对利用主题特征进行作者识别有优化效果，以期结合文体风格特征后能进一步提高作者身份识别的准确性和全面性

1 国内外研究现状

作者识别是指以文本内容和文本属性为依据，抽取出不同作者在文本中所体现的不同特征，进而识别出文本作者[1-3]的研究，而作者特征可以从反映行文风格的文体风格特征和反映文本内容的主题特征两个方向进行体现。文体风格特征表现了作者个人在写作活动中的言语特征，是作者个人风格不自觉的深刻反映，并且这些特征又可以在一定程度上通过数量特征来进行刻画[4]。主题特征则是作者在文章中通过各种材料所表达的中心意思，它渗透、贯穿于文章的全部内容，体现着作者写作的主要意图[5]。

利用文体风格特征进行作者识别的起源最早可追溯到 1887 年 T. C. Mendenhall[6]对戏剧作品文体特征的研究，其研究是使用词汇构建词谱并描绘特征曲线，为莎士比亚戏剧的作者归属争议提供了新的论据，进一步的，研究[7]中被提出使用功能词等特殊词汇，令使用词汇进行作者识别更加精确和有效；De Vel[8]等人则将标点符号等符号特征作为区分不同邮件作者的有效特征，选取的特征在聚合和多主题作者分类识别上都有很好的效果；Keselj 等[9]则提出一种通过计算和比较字符 N-gram 频率识别作者的方法，研究者同时使用该方法在几种不同语言中均进行了作者识别验证，证明了 N-gram 的语言无关性。国内祁瑞华[10-11]团队则是从综合利用文本特征进行作者识别的角度从字符层面、词汇层面、句法层面和结构层面选取特征，建立多层面文体风格特征模型，不仅实现了社交文本作者识别的研究，多特征的选取及在作者识别中的可行性在研究中也得到了验证。

而利用主题特征进行作者识别的研究在早期很少出现，因为主题特征往往反映的是文本的内容，而文本内容在不同体裁，不同情景下很难做到统一，但是仍有研究表明主题特征作为文体特征的补充对于作者识别有积极意义，如 Finn[12]等人通过研究文档分类与文档主题的关系，就发现以同一主题下文档类型分类容易得到较好效果，这说明主题特征对于补充其他特征用于文本分类有一定积极作用。同时，随着主题模型的发展，更多的研究者也开始将其应用于作者识别中，具有代表性的就是 Savoy[13]进行的相关研究，其利用 LDA 分别生成每个作者所有文档的主题模型、待测试文档主题模型，然后计算主题相似度来进行作者归属和识别，而与本研究比较相似的研究是 Waheed Anwar[14]等人提出的实验验证，其利用余弦相似度和 LDA 方法来衡量文本文档向量的相似度，最终达到作者识别的目的，而其在构建的包含 6000 篇文章文档的数据集上进行实验得到的结果表明，该方法优于其他用于作者归属的算法。

2014 年，Y. Nie[15]等人提出，因为人的精力有限，社交网络的使用者围绕的兴趣也有限，这些兴趣中，既有核心兴趣，也存在暂时的边际兴趣。而所谓核心兴趣是指相对稳定的，在短期内不会改变，且在作者发布文本中有较好体现的兴趣，而体现在文本中即是核心主题。2016 年，Shouzhong[16]等人提出利用 Textrank 结合 TF-IDF 对社交网络文本核心主题进行识别，并将其应用于微博数据中。其通过为每个类别分配权重并计算关键字的权重的方法对每个关键字的排名进行评分从而识别出核心主题。

上述针对主题特征和核心主题特征的研究为进一步优化使用主题特征进行作者识别提供了新的思考方向。基于此，本研究以已有的利用 LDA 主题模型进行作者识别的相关研究为技术路线支持[13]，首先验证主题特征在中文社交媒体文本作者识别中具有研究意义，同时从提高主题特征质量的思路出发，以提高作者识别效率为最终目的，提出了使用 LDA 主题模型结合基于 word2vec 的核心主题筛选策略构建候选作者的核心主题识别筛选模型，最终完成科研作者的社交媒体文本的作者识别验证实验。

2 主要技术路线

本研究对科研人员在社交网站上发表的匿名文本进行作者身份识别，而科研人员通常都有自己关注的科研领域，其在社交网站上发布的社交文本信息也会对其科研领域主题有所体现，所以本研究将科研人员的科研领域主题作为核心主题。而每个科研人员的科研主题特征可以从该科研人员发表的学术论文中获得，科研人员社交主题特征可以通过实名社交网站信息获得，所以可以利用学术文本中获得的主题特征对社交网络中的主题特征进行筛选，最终得到本研究的核心主题。

本研究提出利用主题特征进行作者识别的方法框架如图 1 所示，可以概括为 3 个主要步骤：

(1) 作者主题特征抽取。分别从两类数据源抽取两类主题特征：利用科技论文数据抽取候选作者的科研主题特征、利用实名制社交网站数据抽取候选作者社交主题特征。利用 LDA 主题模型分别获得作者学术文本和社交网络文本的主题集合。

(2) 生成作者主题特征模型。识别候选作者核心主题，将科研主题特征与社交主题特征合并，生成候选作者主题特征模型。

(3) 计算待识别文本与作者的相似度。分别计算待识别文本主题特征与各个候选作者主题特征相似度得到最相似的作者作为待识别文本最终的作者识别结果。

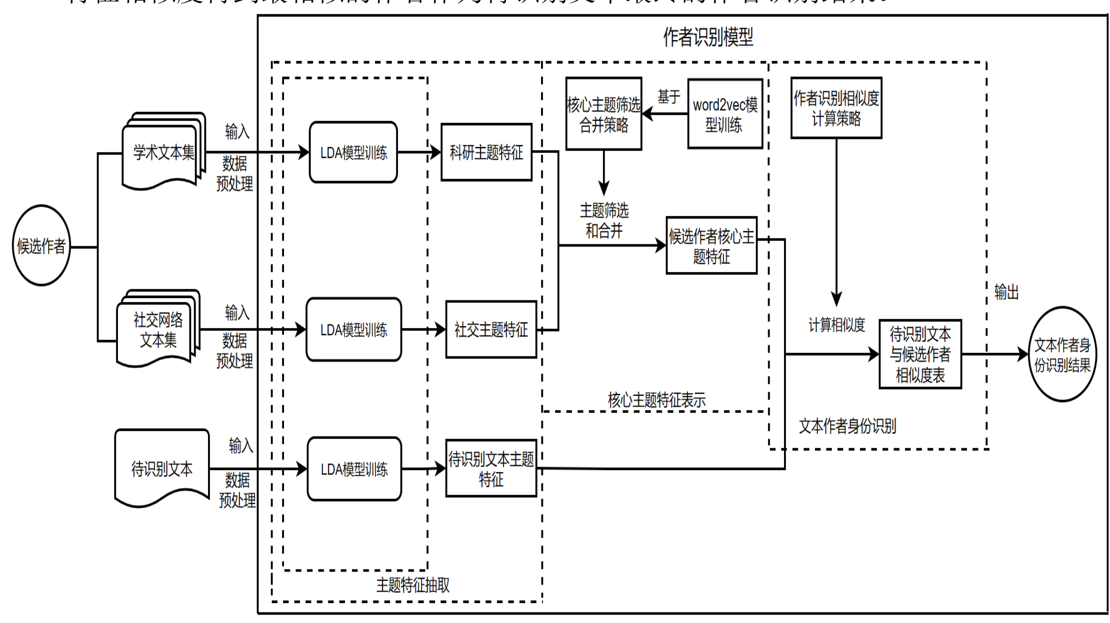


图 1 流程框架

2.1 作者主题识别抽取

对文本的作者身份识别，首选需要获取候选作者的特征，基于 LDA 对主题的抽取能力，LDA 模型经常应用与主题抽取的相关研究中[17]，所以本研究选择采用 LDA 主题模型的方

法对候选作者的主题特征进行识别和表示。

LDA 模型是一种概率主题模型，其基于假设：文档是由若干个隐含主题构成，而这些主题是由文本中若干个特定词汇构成，忽略文档中的句法结构和词语出现的先后顺序[18]。LDA 主题模型由参数(α，β)确定，α 反映了文档集合中隐含主题间的相对强弱，β 刻画所有隐含主题自身的概率分布[19]，从 Dirichlet 分布 α 中取样生成文档-主题分布 θ，从 Dirichlet 分布 β 中取样生成主题-词语分布 φ。

在本研究中，将每个候选作者的文本归为两类文档集，然后分别对这两个文档集进行主题抽取，能够得到每个文档集中每篇文本的文本-主题概率分布，和每个文档集中抽取的主题-词语分布，将每篇文本对应的主题分布概率进行平均，就得到了文本集综合文档-主题概率分布。利用 LDA 从一个文档集抽取出的一个主题集合可以表示为：

其中 T 是主题，由主题词和每个主题词的权重（对主题的贡献度）组成，P 为主题分布概率，即对每篇文章中对应主题的分布概率计算平均值得到的，k 为主题集合 H 中的主题个数。针对 T 展开可以表示为：

其中 T 为主题，由主题词 W 组成，m 为该主题词在主题中的分布概率，c 为主题中包含的主题词的个数。

主题抽取过程最后每个候选作者可以得到两个主题集合：在学术文本中得到的主题矩阵 H_l，代表的是候选作者的科研主题特征，在社交网络文本中得到的矩阵 H_s，代表的是候选作者的社交主题特征。

2.2 核心主题特征计算

2.2.1 核心主题计算启发式规则

本研究认为科研作者的科研主题特征是其核心主题，这些主题特征在其社交主题中可能也会有所体现，所以核心主题筛选的最终目的是找到候选作者社交主题中涉及的核心主题即科研主题。该过程中需要解决的问题主要是如何在候选作者的社交主题特征中找到与其科研主题特征相似的主题以及该通过什么样的手段使其在作者识别中起更重要的作用。针对上述问题，本研究基于以下启发式规则进行策略制定：

(1) 利用 LDA 主题模型抽取出的主题特征由主题词汇组成，两个主题之间的相似度可以通过主题词汇之间的相似度进行衡量。

(2) 利用 LDA 主题模型抽取主题特征，可以获得文档-主题的概率分布和主题-词汇的概率分布，主题的分布概率越大，说明该主题对文档的内容贡献越大，词汇的分布概率越大，说明该词汇对主题的贡献越大。

基于上述的启发式规则，在这一过程中可以尝试利用工具或者模型计算科研主题和社交主题之间的相似度来发现识别核心主题特征，并利用主题和主题词的分布概率来表示主题在识别中的重要性，从而达到核心主题识别筛选的效果。word2vec 利用深度学习的思想，可以从大规模的文本数据中自动学习数据的本质信息[20]，从而通过计算主题词汇之间的相似度达到计算主题相似度的目的。

2.2.2 核心主题特征识别计算

(1) 工具训练

word2vec 模型在给定的语料库上训练 CBOW 和 Skip-Gram 两种模型，然后输出得到所有出现在语料库上的单词的词向量表示[21]。基于得到的单词词向量，可以计算词与词之间的关系，如词语相似性等，从而可以定义主题的相似度，进而计算主题集也就是作者和待识别文本主题特征的相似度，最终达到作者识别的目的。

本文采用开源的 Word2vec 工具，将候选作者的两类文本结合腾讯词向量作为训练数据，用 Skip-gram 模型对训练数据进行训练，得到训练数据中每个词的词向量。表 1 给出了 word2vec 的参数含义及选择，其中 cbow 非 0 时对低频词敏感，size 则是输出词向量的维数，即神经网络的隐藏层的单元数，其取值太小会导致词映射因为冲突而影响结果，值太大则会耗内存并使算法计算变慢，大的 size 需要更多的训练数据，但是效果会更好[22]。参数值的选择是根据已有的研究选定的[23]。表 2 给出了按照相似度排列的词向量训练结果示例。

表 1 Word2vec 参数设置情况

超参数	参数说明	设置
size	词向量的维数	40
window	上下文窗口的大小	20
min-count	词语出现的最小阈值	1
cbow	是否使用 cbow 模型（0 为使用）	3
worker	计算核心	4

表 2 词向量训练结果示例

词语	相近词	相似度	词语	相近词	相似度
计算机	网络 硬件 信息 专业计算机 计算机基础 计算机专业 操作系统 电脑	0.8438003 0.8386404 0.8251010 0.8053851 0.8025305 0.7992088 0.7953601 0.7944639	互联网	因特网 互连网 英特网 互联网络 网际网络 互联网通讯 移动互联网	0.8794277 0.8775173 0.8645825 0.8616846 0.8224803 0.8220754 0.8047673

(2) 核心主题筛选与合并策略

利用候选作者科研主题对其社交主题进行筛选与合并的具体策略可以分为以下几个步骤：

1) 针对有的主题本身的分布概率就很低，能够代表该文本主要内容的概率较低的问题，为了避免主题特征的冗余，首先利用主题在主题矩阵中的分布概率对主题矩阵 H_s 主题特征进行初步筛选。
2) 初筛过后，需要利用相似度计算找到矩阵 H_s 与矩阵 H_l 中相似的主题，将其识别出来，赋予更高的识别权重。因为学术文本主题集合由主题组成，主题由主题词组成，所以此步骤中需要对三个相似度计算进行定义：

   其中 T 为主题，由主题词 W 组成，P 为该主题在矩阵中的分布概率，m 为该主题词在主题中的分布概率。k 为主题矩阵 H 中的主题个数，z 为主题中包含的主题词的个数。

   定义词汇相似度：

   （1）

   定义主题之间的相似度 $sim(T_1,T_2)$：词汇相似度的加权平均，权重是词汇组成主题的概率：

   （2）

   定义主题矩阵（主题集）之间的相似度 $sim(H_l,H_s)$：主题相似度的加权平均，权重是该主题的分布概率:

   （3）

3) 合并规则

筛选合并规则的整体思想是利用上述定义的加权相似度计算的方法，找到每个社交主题特征最相似的科研主题特征，根据阈值判断是否增加其权重将其作为识别过程中的核心主题。同时针对社交主题特征中的主题词，同样利用相同的方法判断其是否相似于科研主题词，并通过阈值判断进行权重重新赋值，而若主题相似主题词不相似，则考虑将科研主题词汇添加到社交主题中用于补充主题特征。具体筛选合并规则如下代码形式所示：

3 实验与结果分析

为了验证主题特征在作者识别中的意义，同时证明本研究核心主题策略算法对利用主题特征进行作者识别的提高效果，本研究的基线实验是利用社交网络文本抽取的原始主题进行作者识别，对比实验是利用学术文本抽取出的主题对社交网络文本主题进行筛选合并后的核心主题进行作者识别；同时因为作者识别任务利用多层次特征是研究发展方向，所以本研究也进行了文体风格特征结合核心主题特征进行作者识别与仅使用文体风格特征进行作者识别的对比实验，用以验证核心主题特征对于文体风格特征的补充作用。

3.1 数据获取和预处理

研究选择计算机领域的 20 位科研人员作为候选作者，利用爬虫分别获取其知网上发布的论文文本以及其在科学网上发布的博客文本作为实验数据集。论文数据共 5612 条，博客数据共 5980 条。利用 jieba 工具进行分词处理，同时对分词结果利用频次和词性等进行筛选，去除人名、停用词、动词、通用词等影响因素，保证主题的表示更具有代表性。随机抽取 20%的科学网文本作为测试集，剩余 80%和全部的论文文本作为训练集进行作者识别模型训练。数据数量如表 3 所示，候选作者学术文本内容和博客文本内容如表 4 所示。

表 3 作者文本数据

作者	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
数量	601	632	813	561	550	627	502	664	731	732	694	803	605	562	563	600	598	571	692	899

表 4 候选作者文本内容示例

	文本内容
学术文本	随着 Internet 的兴起及新的计算模式 (如面向服务的计算和普适计算等)的出现，越来越多的系统以服务组合的方式构建。服务是一种独立的计算组件，分布在网络中的各个设备，通过彼此协作提供服务。……
博客文本	看到一篇文章谈到普适计算与云计算的区别,该文认为云计算是一个可商业实现的平台,它是包含于普适计算当中。换句话说,普适计算的概念更为广泛。本人较认同该观点，但我认为普适计算是提出了一种新的计算模式，目的还是更广泛地资源融合,以及相关技术融合;当然也产生了很多挑战。……

3.2 实验设置

（1） 主题特征抽取

使用 LDA 主题模型获取文本主题，采取开源的 Gibbs 为采样工具，其参数设置如下：模型参数，α、β 分别设为 50/T 和 0.1[24]。至于主题数的选择，研究在对每个候选作者的社交网络文本和学术文本进行 LDA 主题抽取时，首先对训练文本利用困惑度选择主题 T 的可取值范围。困惑度如图 2 所示。

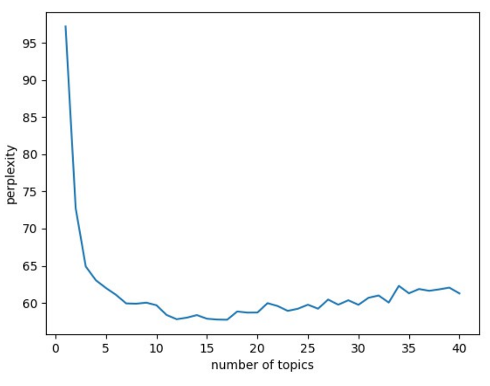


图 2 困惑度曲线图

因为不同的实验语料最佳的主题数是不同的，为了保证实验主题选择的一致性，研究进一步计算了 T 的取值范围上的作者识别效果，最终选择 T=15 作为主题数。表 5 是候选作者核心主题识别计算

表 5 作者 3 的社交网络文本主题分布

主题	主题词	P
T_1	'0.082*互联网','0.030*大脑', '0.012*人类','0.011*网络','0.010*进化' ,'0.009*系统','0.009*结构','0.008*社交','0.008*人脑','0.008*虚拟','0.007*神经系统','0.007*数据','0.006*神经学' ,'0.006*神经','0.006*功能'	0.28120354
T_2	'0.035*威客','0.030*互联网','0.012*知识','0.011*模式','0.009*理论','0.008*科学','0.006*发展','0.006*价值','0.005*智慧','0.004*工作','0.004*功能' '0.004*博客','0.004*进化论','0.004*人类','0.004*进化'	0.19857088
T_3	'0.044*互联网','0.024*智慧','0.019*大脑','0.016*智能','0.012*系统','0.012*反射弧','0.011*人工智能','0.011*人类','0.010*云脑','0.010*神经系统','0.010*进化','0.010*建设','0.009*社会','0.009*架构','0.008*数据'	0.05575676
T_4	'0.002*电平','0.002*电容','0.001*电压','0.001*开关','0.001*拓扑','0.001*输出','0.001*逆变器','0.001*单元','0.001*电流','0.001*矢量','0.001*调制','0.001*载波','0.001*三相','0.001*状态','0.001*共模'	0.11438579
T_5	'0.053*大脑', '0.044*互联网', '0.014*智能', '0.013*人类', '0.010*世界' ,'0.009*发展','0.009*技术', '0.009*系统', '0.009*建设', '0.009*科技','0.007*进化' ,'0.007*模型','0.006*智慧', '0.006*数据','0.005*信息'	0.25008285

（2） 核心主题筛选策略阈值设置

通过计算候选作者学术文本主题矩阵和社交网络文本主题矩阵的相似度，得到平均值 0.0018；通过计算候选作者学术文本主题和社交文本主题的相似度，得到平均值 0.2375。以此为基准设置阈值和系数的优化实验，通过实验迭代，选择 θ_1=0.001、θ_2=0.25 作为较优阈值。而 γ_1、γ_2、ω_1 和 ω_2 的取值，本研究做了迭代实验，结果分别如图 3 和图 4 所示。

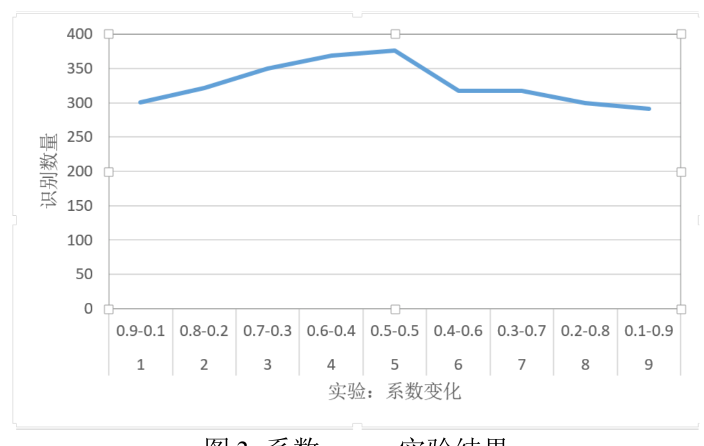


图 3 系数 γ_1、γ_2 实验结果

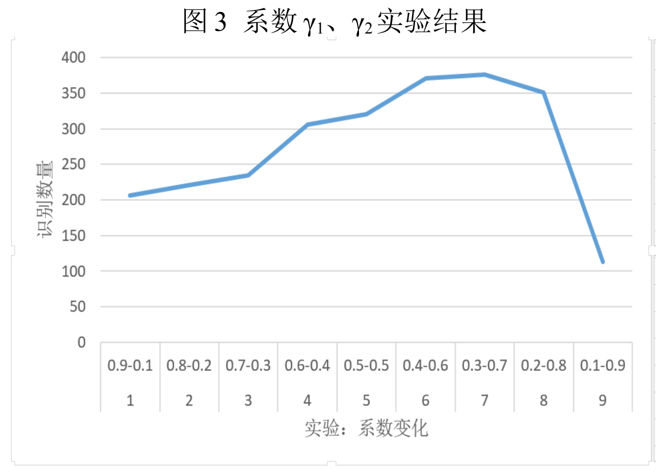


图 4 系数 ω_1、ω_2 实验结果

所以，经过实验优化，选择系数 $\gamma_1=\gamma_2=0.5$，$\omega_1=0.33$，$\omega_2=0.66$。

（3） 评价指标

实验评估方法采用精确率（Precision）、召回率（Recall）和 F1 测试值。可以假设：A 为表示判断为作者 S 且判别正确的文本个数，B 表示判断为写作风格 S 但判别错误 的文本个数，C 表示判断为不属于作者 S 且判别错误的文本个数，D 表示判断为不属于作者 S 且判别正确的文本个数，则我们可以得到指标的计算公式。

（4） （5）

（6）

3.3 实验结果分析

(1) 核心主题特征与主题特征的实验效果对比

表 6 展示了部分代表性作者利用主题特征和核心主题特征结合文体风格特征进行作者识别的结果对比。

表 6 主题特征与核心主题特征进行作者识别结果对比

特征	作者	P	R	F1
主题特征	作者 1	0.8193	0.8000	0.8095
	作者 2	0.8571	0.5455	0.6667
	作者 3	0.4118	0.8750	0.5600
	作者 4	0.2333	0.8750	0.3684
	作者 5	1.0000	1.0000	1.0000
	作者 6	0.5227	0.3965	0.4510
	……		……	
	综合(20 名作者)	0.6674	0.6609	0.6886
核心主题特征	作者 1	0.9512	0.9176	0.9341

	作者 2	0.8000	0.7273	0.7619
	作者 3	0.5000	0.8750	0.6364
	作者 4	1.0000	0.8750	0.9333
	作者 5	0.9091	1.0000	0.9524
	作者 6	0.6774	0.3621	0.4719
	……	……	……	……
	综合（20 名作者）	0.7837	0.8276	0.8521

分析表 6 可以发现：

①综合来看利用核心主题特征进行作者识别的三个衡量指标都有所提高，这说明了利用学术主题特征对社交网络文本主题特征进行筛选合并得到的核心主题特征应用于作者识别，能一定程度提高识别的准确率，有一定的优化作用。

②具体到候选作者，可以看到大部分的作者的识别效果能得到一定的提升，但是以作者 6 为代表的候选作者 P 指标和 F1 指标均有所下降，分析数据可以发现该利用核心主题的方法针对在社交网络中主题较为集中，且有较大比例涉及到其在学术中的研究领域的作者更有效果；而针对在社交网络文本中不涉及或者少量涉及学术领域的作者，该方法取得的优化效果较小。

(2) 核心主题特征对文体风格特征的补充验证实验

根据已有的研究，仅使用一种特征进行作者识别的效果是不突出的，多层次特征结合使用才是作者识别的发展方向。所以为了验证本研究的核心主题特征对于文体风格特征有补充作用，对于结合其他特征进行作者识别也有进一步的研究前景，下面进行核心主题特征对文体风格特征的补充验证实验。

本研究选择的文体风格特征是 N-gram 特征，它可以捕捉到作者风格的细微差别，包括由词汇、上下文、标点符号以及大小写变动所带来的差别[25]，表示方便，识别效率较高。因为 N-gram 特征的抽取和使用已经较为成熟，所以下面仅阐述结合其在实验中的使用。研究利用两种特征分别计算待识别文本与候选作者的相似度，然后对相似度进行加权分析，相似度最高的作者作为最终的识别结果。加权系数经过多次交叉实验，其他系数和影响因素不变的情况下，选择文体风格特征系数为 0.82，主题特征系数为 0.18 时识别的文本数最多，效果最好，故以此为特征系数。

本实验用 CountVectorizer 方法，设置阈值为 min_df=2，基于此构建作者的 N-gram 特征向量。表 8 是作者 2 的部分 N-gram 特征。

表 8 作者 N-gram 特征示例

作者	N-gram
作者 2	{('软件工程', '软件工程'): 2, ('软件工程', '专业'): 2, ('专业', '必修'): 2, ('必修', '专业'): 2, ('专业', '基础课'): 2, ('涉及', '内容'): 2, ('包含', '软件'): 2, ('软件', '生命周期'): 2, ('生命周期', '阶段'): 2, ('阶段', '需要'): 2, ('需要', '知识'): 2, ('一门', '概论'): 2, ('概论', '性质'): 2, ('性质', '课程'): 2}

表 9 展示了仅使用文体风格特征识别的和结合文体风格特征与核心主题特征识别的结果对比。

表 9 实验特征组合识别结果

特征	作者	F1	R	P
N-gram 特征	作者 1	0.8000	0.6824	0.9667
	作者 2	0.5333	0.7273	0.4211
	作者 3	0.4242	0.8750	0.2800
	作者 4	0.4827	0.8750	0.3333
	作者 5	0.9524	1.0000	0.9091
	作者 6	0.4800	0.4138	0.5714

	……	……	……	……
	综合（20 名作者）	0.6326	0.5690	0.6336
N-gram 特征+核心主题特征	作者 1	0.9512	0.9176	0.9341
	作者 2	0.8000	0.7272	0.7619
	作者 3	0.5000	0.8750	0.6364
	作者 4	1.0000	0.8750	0.9333
	作者 5	0.9091	1.0000	0.9524
	作者 6	0.6774	0.3621	0.4719
	……	……	……	……
	综合（20 名作者）	0.7837	0.8276	0.8521

分析表 9，可以发现：

①从综合结果来看，利用核心主题结合文体风格特征作者识别的效果要优于仅使用 N-gram 特征进行识别，这说明在该实验集上，核心主题特征的使用对作者识别有积极作用。

②具体到每个候选作者，可以看到，多数的作者主题特征的识别效果是积极的，这也充分论证了科研人员的领域主题能一定程度上成为该作者标签特征，这是具有个人性的特征。而针对作者 5 为代表的作者，其 F1 值降低，作者 6 为代表的作者，其召回率和精准率均降低，则认为主题特征未起到积极效果，笔者分析其文本认为这与其所关注的领域较为宽泛且学术领域与科研文本中的主题相差较大，以至于本研究的核心主题筛选合并策略未起到较大作用，而添加主题特征作为识别特征相当于增加了干扰项，导致识别准确率下降。而针对这一现象，后续可以通过分步式结合两种特征的方法进行改善，如先利用主题特征进行作者识别，给出相似的几个候选作者，缩小候选作者数量，然后进一步通过 N-gram 特征得到最相似的候选作者作为识别结果。

③分析不同候选作者的识别效果，发现训练训练语料的体量也会影响主题特征在作者识别上的应用效果，在目前实验中，训练数据越多，抽取出的主题特征越具有代表性，识别的准确性也越高。候选针对该影响，可以继续通过控制实验数据大小进行对比实验。

(3) 结论

通过对结果的分析可以看出，在核心主题特征对于主题特征的优化方面，可以发现使用筛选得到的核心主题进行作者特征表示并用于作者识别的效果优于仅利用原始抽取出的主题特征，这有效证明，针对科研人员的社交网络文本的作者识别，利用其在学术文本中所体现的领域主题对其社交网络文本主题特征进行筛选和合并，能够进一步优化主题特征筛选并给予识别作用更大的特征更高的权重，从而提高作者识别的准确率，进而优化作者识别效果；在核心主题特征结合其他特征在作者识别中的应用效果方面，核心主题特征能够有效地提高仅利用 N-gram 特征作者识别的效果，这证明在本研究的实验语料上，核心主题特征对于结合文体风格特征用于作者识别有一定的积极意义。

4 总结展望

本研究重点研究了结合学术文本对利用主题特征进行作者识别的可行性，以及其的优化策略。通过实验验证，结果显示优化策略是有效的，在此进行总结并讨论未来可以继续优化的内容：

(1) 本研究重心在主题特征的使用优化方向，可以发现利用学术文本的领域主题对候选作者的社交网络文本主题进行筛选得到的核心主题特征有提高利用主题特征作者识别效果的作用；研究也尝试结合 N-gram 特征和核心主题特征，同时对比仅使用 N-gram 特征，结合主题特征对作者识别也有一定程度上的提升。进一步研究可以考虑从其他文体风格特征出发，或者结合多层次文体风格特征进行应用研究。

(2) 本研究在对待识别文本进行作者识别实验时，使用 LDA 主题模型作为主题抽取的方法，而随着相关研究的发展，其他主题模型或者其他主题抽取方式或许较 LDA 主题模型能取得不同的效果。所以针对主题获取这一步骤，后续研究可以尝试从主题获取方法入手，尝试进一步优化识别效果。

(3) 本研究目前仅考察利用文本的内容信息进行作者识别，未来随着网站文本属性或者用户属性的完善，亦可以考虑借助社交网络相邻用户的文本信息和属性信息进行特征抽取和核心主题选择，相信能进一步提高作者识别效果。

参考文献

[1] Kalgutkar V, Kaur R, Gonzalez H, et al. Code authorship attribution: Methods and challenges[J]. ACM Computing Surveys (CSUR), 2019, 52(1): 1-36.

[2] Alrabaee S, Debbabi M, Wang L. CPA: Accurate Cross-Platform Binary Authorship Characterization Using LDA[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3051-3066.

[3] I. Maglogiannis, L. Iliadis ,E. Pimenidis. Artificial Intelligence Applications and Innovations[J]. IFIP Advances in Information and Communication Technology, 2020,583:55–266.

[4] 刘颖,肖天久.金庸与古龙小说计量风格学研究[J].清华大学学报(哲学社会科学版),2014,29(05):135-147,179.

[5] 百度百科.主题.[EB/OL]. [2022-9-5].https://baike.baidu.com/item/主题/2894698.

[6] Mendenhall T C. The characteristic curves of composition[J].Science,1887(214S):237-246.

[7] Hoover, D.L. Another Perspective on Vocabulary Richness[J].Computers and the Humanities,2003,37:151–178.

[8] De Vel O, Anderson A, Corney M, et al. Mining e-mail content for author identification forensics[J]. ACM SIGMOD Record,2001,30(4):55-64.

[9] Keselj V, Peng FC,Cercone N, Thomas C.N-gram based author profiles for authorship attribution[C]//Pacific Association for Computational Linguistics. Halifax, Canada: PACL, 2003:255−264.

[10] 祁瑞华,杨德礼,郭旭,刘彩虹.基于多层面文体特征的博客作者身份识别研究[J].情报学报,2015,34(06):628-634.

[11] 祁瑞华,郭旭,刘彩虹.中文微博作者身份识别研究[J].情报学报,2017,36(01):72-78.

[12] Finn A, Kushmerick N. Learning to classify documents according to genre[J]. Journal of the American Society for Information Science and Technology, 2006, 57(11): 1506-1518.

[13] Savoy J. Authorship attribution based on a probabilistic topic model[J]. Information Processing & Management, 2013, 49(1): 341-354.

[14] [70] W. Anwar, I. S. Bajwa, M. A. Choudhary and S. Ramzan.An Empirical Study on Forensic Analysis of Urdu Text Using LDA-Based Authorship Attribution[J]. IEEE Access,2019(7): 3224-3234.

[15] Y. Nie, J. Huang, A. Li, B. Zhou. Identifying users based on behavioral-modeling across social media sites[J].Web Technologies and Applications, 2014,8709:48–55.

[16] Tu Shouzhong, Huang Minlie.Mining microblog user interests based on TextRank with TF-IDF factor[J].The Journal of China Universities of Posts and Telecommunications,2016,23(5): 40-46.

[17] 陈思含.基于微博的多特征情感分析方法研究[D].吉林大学,2021.

[18] 姚全珠,宋志理,彭程.基于 LDA 模型的文本分类研究[J].计算机工程与应用, 2011, 47(13): 150-153.

[19] 王振振,何明,杜永萍.基于 LDA 主题模型的文本相似度计算[J].计算机科学, 2013,40(12):229-232.

[20] 马思丹,刘东苏.基于加权 Word2vec 的文本分类方法研究[J].情报科 19,37(11):38-42.

[21] 李晓,解辉,李立杰.基于 Word2vec 的句子语义相似度计算研究[J].计算机学,2017, 44(09): 256-260.

[22] 你好星期一.Word2vec 参数[EB/OL]. [2022-12-13]. https://blog.csdn.net/DL_Iris/article/details/119175496.

[23] 张谦,高章敏,刘嘉勇.基于 Word2vec 的微博短文本分类研究[J].信息网络安全,2017(01):57-62.

[24] 唐晓波,祝黎,谢力.基于主题的微博二级好友推荐模型研究[J].图书情报工作,2014, 58(09):105-113.

[25] Johnson A, Wright D. Identifying idiolect in forensic authorship attribution: An N-gram text bite approach[J].Language and Law, 2014,1(1):37−69.